Cross-validatory Credibility in Prediction

(Preliminary Report)

by

David V. Hinkley*
University of Minnesota
Technical Report No. 240

* Research partially supported by U.S. Army grant DAHC04-74-G-0216

Cross-validatory Credibility in Prediction

(Preliminary Report)

By David Hinkley

0. Abstract

This report contains preliminary ideas on how to obtain credibility intervals for future observations in a stable observed process without model assumptions, using the notion of cross-validation or predictive sample reuse. Some elementary examples are given. The results provide some alternative justification for procedures derived through parametric models.

1. Introduction

We are concerned with prediction of future observations in a stable process on which data are available. For simplicity we consider only those cases where observable variables are essentially exchangeable, which in parametric modelling would usually be described as follows: $Y_1,\ldots,Y_{n+1}$ are i.i.d. observable variables with p.d.f. $f(y;\theta)$, $\theta \in \Omega$. Given data values $y_1,\ldots,y_n$ we are required to make some predictive statement about the as-yet unseen value y_{n+1}.

A standard statistical problem, given $\{f(y;\theta),\theta \in \Omega\}$, is to produce a "good" point predictor $\hat{y}_{n+1} = g_{n+1}(y_1,\ldots,y_n)$, say, where "good" usually means "low mean-squared error." In certain situations one would find a point predictor of little value, more informative being one or more probability-type statements such as $\mathrm{pr}(Y_{n+1} \in A) = p$. It is this kind of inference about the future which interests us here.

A genuine probability statement of the type alluded to above is available only in the Bayesian framework. Without the extra structure of a prior distribution over Ω we can at best obtain statements of the form

$$\mathrm{pr}(Y_{n+1} \in A(Y_1,\ldots,Y_n)) = p, \tag{1.1}$$

which implies a frequency interpretation to the statement "$Y_{n+1} \in A(y_1,\ldots,y_n)$" valid over repetitions of $(y_1,\ldots,y_{n+1})$. The confidence p attached to "$Y_{n+1} \in A(y_1,\ldots,y_n)$" is not a probability if $y_1,\ldots,y_n$ is fixed, although under certain models p is a (fiducial) probability within a more relevant series of hypothetical repetitions than implied by (1.1); see Fraser (1961). We note in passing that even statements such as (1.1) cannot be obtained exactly in many parametric models.

The common link between Bayesian, fiducial and sampling-theory frameworks is the parametric model $\{f(y;\theta),\ \theta \in \Omega\}$, which is often only tentatively asserted, and is always rigidly followed. (It is interesting that estimating θ corresponds to predicting the indefinite future $Y_{n+1}, Y_{n+2},\ldots$, whereas usually we only wish to predict a finite future.) In any event, if the model <u>is</u> tentative, it certainly makes sense to compare a prediction based on $y_{i_1},\ldots,y_{i_{n-1}}$ with the realization y_{i_n}, for some or all permutations $(i_1,\ldots,i_n)$ of $(1,\ldots,n)$. This notion of assessing internal consistency or validity is at the heart of cross-validation and jackknife procedures. Recently, Stone (1974) and Geisser (1975) have shown how <u>point</u> predictors $g_{n+1}(y_1,\ldots,y_n)$ can be obtained using only a family of predictive prescriptions.

$$\mathcal{G} = \{g_{n+1}(y_1,\ldots,y_n,\alpha);\ n = 1,2,\ldots;\ \alpha \in A\}$$

and the cross-validation mechanism with suitable discrepancy measure. That is, no hypothetical probability model $\{f(y;\theta),\theta \in \Omega\}$ is assumed.

In Section 2 we present a method of generating credibility (not true probability) statements about y_{n+1} by applying cross-validation (sub-sampling) ideas. This is based on a notion of predictive likelihood, and is unconnected with point prediction.

2. <u>Predictive sufficiency and cross-validation distributions</u>

To motivate the first cross-validation method we return briefly to

parametric prediction. Suppose that Y and θ are one-dimensional, the p.d.f. of Y given θ being

$$f(y;\theta) = \exp\{\theta b(y) + c(\theta) + d(y)\} , \tag{2.1}$$

and now suppose $Y_1,\ldots,Y_{n+m}$ are i.i.d. with p.d.f. (2.1). Then one natural analog of the likelihood for θ is the predictive likelihood for $Y_{n+1},\ldots,Y_{n+m}$ defined by

$$\text{plik}(y_{n+1},\ldots,y_{n+m}|y_1,\ldots,y_n) \propto f_{S_n|S_{n+m}}(s(y_1,\ldots,y_n)|s(y_1,\ldots,y_{n+m})), \tag{2.2}$$

where $S_k = s(Y_1,\ldots,Y_k) = \sum_{j=1}^{k} b(Y_j)$ is the sufficient reduction of $(Y_1,\ldots,Y_k)$, $k \geq 1$. In effect, plik( ) measures the credibility that would be attached to the <u>already observed</u> value $s_n = s(y_1,\ldots,y_n)$ if in fact the sequence continued on to yield $(y_1,\ldots,y_n,y_{n+1},\ldots,y_{n+m})$. This definition, proposed independently by Hinkley (1975) and Lauritzen (1975), has the important property of removing θ. Note that given s_{n+m}, $y_{n+1},\ldots,y_{n+m}$ have a known exchangeable distribution independent of θ which is subsumed in the constant of proportionality in (2.2). It is worth pointing out that plik( ) is a Bayes factor regardless of the prior distribution on θ. For further details and properties of (2.2) we refer to Hinkley (1975).

From the model-free viewpoint, the key ideas in the above are (a) the summarizing statistic s(.) for any data sequence, and (b) the conditional frequency (2.2) as a measure of credibility for $y_{n+1},\ldots,y_{n+m}$. If we choose <u>any</u> relevant summary statistic s(.), we can in principle generate frequencies such as plik( ) inside the data $y_1,\ldots,y_n$ by sub-sampling. To illustrate the general idea we examine in detail the simple case of binary data.

<u>Example 1 Homogeneous binary trials</u>

Let $y_1,\ldots,y_n$ be binary (zero-one) variables, n fixed, and suppose their order can be assumed irrelevant. As a summary statistic take $s_k = \sum_{j=1}^{k} y_j$,

and suppose $s_n = r \geq 1$. Consider prediction of y_{n+1} in the sense of credibility (2.2), for which we require

$$g_{n+1}(z|s_n) = pr(S_n = s_n | S_{n+1} = s_n+z). \tag{2.3}$$

Since the argument of S_{n+1} is to vary, we examine all pairs (s_{n-2}, S_{n-1}) generated by data subsamples, i.e. pairs

$$s_{n-2;i,j} = \sum_{k \neq i,j} y_k \quad , \quad s_{n-1;i} = \sum_{k \neq i} y_k \quad .$$

The following frequency table is obtained

s_{n-2} \ s_{n-1}	r-1	r	total
r-2	r(r-1)	0	r(r-1)
r-1	r(n-r)	r(n-r)	2r(n-r)
r	0	(n-r)(n-r-1)	(n-r)(n-r-1)
total	r(n-1)	(n-r)(n-1)	n(n-1)

Table 1. Subsample frequencies of (s_{n-2}, s_{n-1}) for binary data with r ones in n observations.

From the table we deduce the empirical probabilities

$$pr(s_{n-2} = r-1 | s_{n-1} = r-1+z) = \left(\frac{n-r}{n-1}\right)^{1-z} \left(\frac{r}{n-1}\right)^{z} \quad (z = 0,1) \tag{2.4}$$

which by comparison with (2.3) gives sample predictive likelihood

$$g_{n+1}(z|r) = \left(\frac{n-r+1}{n+1}\right)^{1-z} \left(\frac{r+1}{n+1}\right)^{z} \quad , \quad (z = 0,1). \tag{2.5}$$

The relative credibilities attached to $y_{n+1} = 0, 1$ are $\frac{n-r+1}{n+2}$ and $\frac{r+1}{n+2}$,

which amounts to weighted combinations of empirical proportions and the uninformed prior proportion $\frac{1}{2}$.

This analysis extends to prediction of $\sum_{j=1}^{m} y_{n+j}$, for which we subsample all pairs (S_{n-2m}, S_{n-m}), implying the restriction $2m < n$. The induced predictive sample likelihood is

$$g_{n+1,n+m}(z|r) = \frac{\binom{n}{r}\binom{m}{z}}{\binom{m+n}{r+z}} \qquad (z = 0, 1, \ldots, m), \tag{2.6}$$

which correspond loosely to binomial probabilities with parameter a weighted conbination of r/n and 1/2. The result (2.6) is exactly that obtained by the parametric binomial model in (2.2).

The preceding analysis is not directly relevant if the data are not exchangeable, for example if $y_1, \ldots, y_n$ were obtained by inverse sampling to the fixed number r of successes. In such a situation most sampling-theory prediction procedures are inoperative, although the predictive likelihood (2.2) can be used. In the present context, if the sampling rule is to stop at r ones, then $y_1, \ldots, y_{n-1}$ are exchangeable given $y_n = 1$. It is not immediately clear how to proceed from this, since the fully-informative prediction must use the event "$y_n = 1$." One method is to obtain the credibility in terms of

s_k = no. of trials to obtain k ones,

i.e. compute the credibility of $s_{r+1} = n+1$ by extrapolation from the sub-sampling table of

$$\text{pr}(s_{r-2} = m_2 \mid s_{r-1} = m_1) \quad .$$

A simple combinatorial calculation shows that this probability is

$$\binom{m_2-1}{r-3} \div \binom{m_1-1}{r-2} \qquad (r-2 \le m_2 \le m_1 \le n-1),$$

defined only for $r \ge 2$. Extrapolating $m_2 \to n$, $r-2 \to r$ and $m_1 \to n+1-z$ would give the extrapolated probability

$$h(z|r,n) = \binom{n-1}{r-1} \div \binom{n+z}{r} \quad , \tag{2.7}$$

with relative credibility

$$h(0|r,n) \div \sum_{z=0}^{\infty} h(z|r,n)$$

at z = 0. This seems to be a rather ad hoc way to proceed, and illustrates the kind of difficulty in any non-Bayesian prediction analysis. Nevertheless, the solution (2.7) would be reasonably acceptable, and would agree closely with (2.5) except in extreme cases.

The preceding example can be generalized directly to multinomial data characterized by cell frequencies. Briefly, if c cells have observed frequencies $r_1,\ldots,r_c$ with $n = \Sigma r_j$ fixed, and if a single observation is to be predicted, then the analog of (2.5) is equivalent to attaching credibility $(r_j+1)/(n+c)$ to the j^{th} cell. We omit details.

In principle the method of subsampling used above can be used with any summary statistic. In general, however, there is a difficulty arising from the fact that data values are often distinct. The multinomial frequency characterization with n cells would not lead to very useful results in such a situation--nor would the sample average summary come to that. Inevitably we would need to introduce a smoothing device at some point in the analysis, and if we stuck close to the multinomial frequency representation we might as well group data to begin with. This can be accomplished in a flexible way using the algorithm proposed by Lindsey (1974), which roughly speaking compares the multinomial likelihoods for several cell widths in an attempt to see when the cell widths alone, rather than the data, change the likelihood.

If one were working with the sample average $\bar{y}_n$ as data summary, one could either use the discretized results of Lindsey's algorithm as input data, or construct the sub-sampling table of $\bar{y}_{n-1;i}$ and $\bar{y}_{n-2;i,j}$ and smooth it so as

to retain major features of the data. The latter method could involve a very large (i.e., $n(n-1) \times n$) preliminary table of zeros and ones, so that the former method would seem preferable. We have not investigated this in any depth.

As a second example of the sample predictive likelihood we take the situation where data are summarized by their maximum.

Example 2 Upper bound statistic

Suppose that $y_1,\ldots,y_n$ are distinct and that we take $s_n = y_{(n,n)}$, where $z_{(m,j)}$ is the j^{th} largest of $z_1,\ldots,z_m$. Thus we are describing the data by the lowest known upper bound. In sub-samples of size n-1 and n-2 the possible values of s_{n-1} and s_{n-2} are respectively $\{y_{(n,n)}, y_{(n,n-1)}\}$ and $\{y_{(n,n)}, y_{(n,n-1)}, y_{(n,n-2)}\}$. It is easy to obtain the following table of sub-sampling frequencies.

s_{n-2} \ s_{n-1}	$y_{(n,n-1)}$	$y_{(n,n)}$	total
$y_{(n,n-2)}$	1	0	1
$y_{(n,n-1)}$	n-2	n-1	2n-3
$y_{(n,n)}$	0	(n-1)(n-2)	(n-1)(n-2)
Total	n-1	$(n-1)^2$	n(n-1)

Table 2. Sub-sampling frequencies of upper-bound statistic.

Therefore, under exchangeability,

$$pr(s_{n-2} = y_{(n,n-1)} | s_{n-1} = y_{(n,n)}) = \frac{1}{n-1} \quad , \tag{2.7}$$

and

$$pr(s_{n-2} = y_{(n,n-1)} | s_{n-1} = y_{(n,n-1)}) = 1 - \frac{1}{n-1} \quad . \tag{2.8}$$

Extrapolating as we did in Example 1 would lead from (2.7-8) to

$$\text{pr}(s_n = y_{(n,n)} | s_{n+1} = y_{n+1}) = 1 - \frac{1}{n+1} \tag{2.9}$$

and

$$\text{pr}(s_n = y_{(n,n)} | s_{n+1} = y_{n+1}) = \frac{1}{n+1} \quad (y_{n+1} > y_{(n,n)}), \tag{2.10}$$

There is some difficulty of interpretation here, but it would be reasonable to infer that credibility (2.9) is attached to the event "$y_{n+1} \leq y_{(n,n)}$", since

$$\{s_{n+1} = y_{(n,n)}\} \equiv \{s_n = y_{(n,n)}\} \cap \{y_{n+1} \leq y_{(n,n)}\} \ .$$

On the face of it, (2.10) gives equal credibility to all values $y_{n+1} > y_{(n,n)}$, with credibility $(n+1)^{-1}$ to the whole set. One is tempted to coalesce this credibility on the value

$$y_{n+1} = y_{(n,n)} + y_{(n,n-1)} - y_{(n,n-2)} \ ,$$

but there is no logical basis for this. Note that for <u>any</u> continuous probability distribution we have the repeated-sampling probability

$$\text{pr}(Y_{n+1} > Y_{(n,n)}) = \frac{1}{n+1} \ .$$

It is important to stress the point that the prediction statements induced here relate to whether or not y_{n+1} exceeds a particular value. This dependence of prediction on data characteristic is quite general. The same is true when s_n is any single order statistic. For example with s_n = median, we find $\text{pr}(y_{n+1} > \text{median}) = \frac{1}{2}$.

Before we leave this preliminary study of sub-sampling, we should note that the method is <u>not</u> the same as computing conditional frequencies

$$\text{pr}(S_{n+1} = s_{n+1} | S_n = s_n)$$

by extrapolation from sub-sampled frequencies

$$\text{pr}(S_{n-1} = s_{n-1} | S_{n-2} = s_{n-2}).$$

For instance, in Example 1 we would have, from Table 1,

$$\mathrm{pr}(S_{n-1} = r - 1 + z \mid S_{n-2} = r - 1) = \tfrac{1}{2} \qquad (z = 0, 1),$$

which is a useless result.

3. Summary

The sub-sampled empirical version of predictive likelihood (2.2) generates results corresponding closely to parametric results in the Bernoulli, multinomial and upper bound examples. In this sense the cross-validation point of view lends support to parametric methods. The results themselves have intuitive appeal. However the method seems to be of very limited applicability. It will be worth investigating examples where Lindsey's (1974) algorithm is used in conjunction with the multinomial results.

The major thrust of cross-validation methods has been in the development of point predictors. Clearly future study of credibility intervals for prediction should include methods of building intervals, or regions, centered on the point predictors, making use of the cross-validatory assessment of those predictors (Stone, 1974; Geisser, 1975).

4. References

Fraser, D.A.S. (1961) The fiducial method and invariance. Biometrika 48, 261-80.

Geisser, S. (1975) The predictive sample reuse method with applications. JASA 70 (to appear)

Hinkley, D. V. (1975) The uncertain future (in preparation)

Lauritzen, S. (1975) Sufficiency, prediction and extreme models. Scand. J. Statist. 2 (to appear)

Lindsey, J. K. (1974) Comparison of probability distributions. J.R.S.S. B 36, 38-47.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. JRSS B, 36 (to appear)